Beyond Gigabit Ethernet: Physical Layer Issues in Future Optical Networks

L B James, A W Moore, R Plumb, M Glick, A Wonfor, I H White, D McAuley, R V Penty

IRC –TR-04-025

September, 2004.

London Communications Symposium

DISCLAIMER: THIS DOCUMENT IS PROVIDED TO YOU "AS IS" WITH NO WARRANTIES WHATSOEVER, INCLUDING ANY WARRANTY OF MERCHANTABILITY NON-INFRINGEMENT, OR FITNESS FOR ANY PARTICULAR PURPOSE. INTEL AND THE AUTHORS OF THIS DOCUMENT DISCLAIM ALL LIABILITY, INCLUDING LIABILITY FOR INFRINGEMENT OF ANY PROPRIETARY RIGHTS, RELATING TO USE OR IMPLEMENTATION OF INFORMATION IN THIS DOCUMENT. THE PROVISION OF THIS DOCUMENT TO YOU DOES NOT PROVIDE YOU WITH ANY LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS

Copyright 2004, Intel Corporation, All rights reserved.

Intel Research Cambridge

Beyond Gigabit Ethernet: Physical Layer Issues in Future Optical Networks

L B James †, A W Moore ‡, R Plumb †, M Glick §, A Wonfor †, I H White †, D McAuley §, R V Penty †

† Centre for Photonic Systems, Department of Engineering, University of Cambridge ‡ Computer Laboratory, University of Cambridge § Intel Research Cambridge

Abstract: This paper presents a study of the errors observed on an optical Gigabit Ethernet link in a state of low receiver power. This condition is increasingly likely as networks become more complex, with longer fibre lengths, optical switching systems and higher data rates. We discover that some octets and sequences of octets have a far higher probability of being received in error than others. This non-uniformity of error in the physical layer may severely affect network performance at higher levels, and should be carefully considered as the next generation of optical networks is developed.

1 Optical networking motivation.

Current work in all areas of networking has led to increasingly complex architectures: our interest is focused upon the field of optical networking, but this is also true in the wireless domain. Our exploration of the robustness of network systems is motivated by the increased demands of these new optical systems. To take advantage of capacity developments offered by optical systems at the short timescales relevant to local area networks, packet switching and burst switching techniques have seen significant investigation. Our project [1,2] is to investigate Optical Packet Switching (OPS) through the construction of a switched optical data path based upon semiconductor optical amplifiers (SOAs). As part of this work we recognise that the need for higher data-rates and designs with larger numbers of optical components are forcing us towards what traditionally have been technical limits.

A significant amount of work exists that is targeted towards the wide-area telecommunications arena. In contrast, we are motivated by an OPS for short link lengths — those comparable with current system-area and local-area networks. The main criteria of our approach are low latency and the avoidance of optical buffering and all optical signal-processing to ensure lowest cost. Our architecture uses high-speed optical switch fabrics for high capacity routing and combines this with wavelength-striping and a separate control channel [2, 3]. Of particular note is the data path between the sending and receiving end-systems: this path is clearly non-trivial with a significant number of devices such as SOAs and wavelength multiplex and de-multiplex units.

There have also been changes in the construction and needs of fibre-based computer networks. In deployments containing longer runs of fibre using large numbers of splitters for measurement and monitoring and active optical devices, the overall system loss may be greater than in today's point-to-point links and the receivers may have to cope with much-lower optical powers. Ethernet in the first mile [4], along with a new generation of switched optical networks, are examples of this trend. In addition, if all other variables are held constant an increase in bandwidth will require a proportional increase in transmitter power. Fibre nonlinearities impose limitations on the maximum optical power able to be used in an optical network. Subsequently, we maintain that a greater understanding of the low-power behaviour of coding schemes will provide invaluable insight for future systems.

We selected 8B/10B coding as the basis for our work, for which initial results are given in James *et al.* [5]. This codec, originally described by Widmer & Franaszek [6], is widely used. It converts 8 bits of data for transmission (ideal for any octet-orientated system) into a 10 bit line code. Although this adds a 25% overhead, 8B/10B has many valuable properties; a transition density of at least 3 per 10 bit code group and a maximum run length of 5 bits for clock recovery, along with virtually no DC spectral component. In addition to being the standard Physical Coding Sublayer (PCS) for Gigabit Ethernet [7], it is used in Fibre Channel, the 800Mbps extensions to the IEEE 1394 / Firewire standard, and will be the basis of coding for the electrical signals of PCI Express.

2 Bit Error Rate versus Packet Error Experiments.

We investigate Gigabit Ethernet on optical fibre (1000BASE-X [7]), under conditions where the received power is sufficiently low as to induce errors in the Ethernet frames. We assume that while the Functional

Redundancy Check (FRC) mechanism within Ethernet is sufficiently strong to catch the errors, the dropped frames and resulting packet loss will result in a significantly higher probability of packet errors than the norm for certain hosts, applications and perhaps users.

In our main test environment an optical attenuator is placed in one direction of a Gigabit Ethernet link. A traffic generator feeds a Fast Ethernet link to an Ethernet switch, and a Gigabit Ethernet link is connected between this switch and a traffic sink and tester. The variable optical attenuator and an optical isolator are placed in the fibre in the direction from the switch to the sink. A packet capture and measurement system is implemented within the traffic sink using an enhanced driver for the SysKonnect SK-9844 network interface card (NIC), which allows application processes to receive error-containing frames that would normally be discarded. As well as purpose-built code for the receiving system we use a special-purpose traffic generator and comparator. Pre-constructed test data in tcpdump-format is transmitted from one or more traffic generators using an adapted version of *tcpfire* [8]. Transmitted frames are compared to their received versions and if they differ, both original and errored frames are stored for octet-by-octet analysis.

Some results presented here are conducted with real network traffic referred to as the *day-trace*. This network traffic was captured from the interconnect between a large research institution and the Internet over the course of two working days. Other traffic tested included *pseudo-random data*, consisting of a sequence of frames of the same number and size as the *day-trace* data, although each is filled with a stream of octets whose values were drawn from a pseudo-random number generator. *Structured test data* consists of a single frame containing repeated octets: 0x00–0xff, to make a frame 1500 octets long. The *low error testframe* consists of 1500 octets of 0xCC data (selected for a low symbol error rate); the *high error testframe* is 1500 octets of 0x34 data (which displays a high symbol error rate).

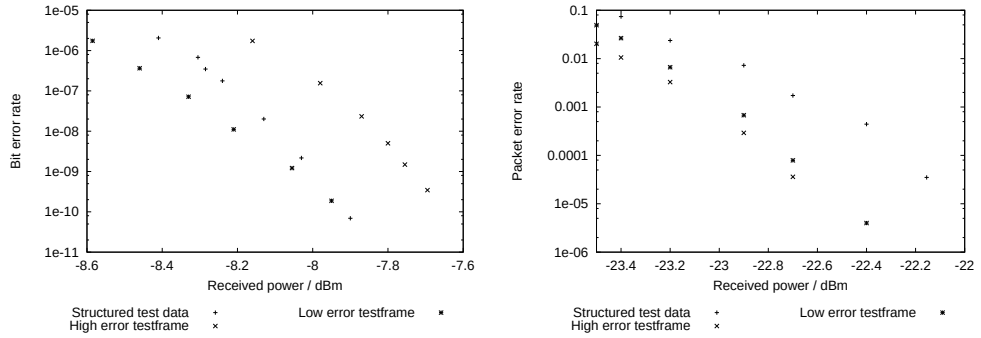
To contrast with these packet and octet error measurements, we also use a Bit Error Rate Test kit (BERT), as would commonly be used in an optical communications laboratory. For the BER measurements presented here, a directly modulated 1548nm laser was used. The optical signal was then subjected to variable attenuation before returning via an Agilent Lightwave (11982A) receiver unit into the BERT (Agilent parts 70841B and 70842B). The BERT was programmed with a series of specially generated bit sequences, each corresponding to a frame of Gigabit Ethernet data encoded as it would be for the line in 8B/10B.

Initial Results. We illustrate how errors are position independent but dependent upon the encoded data [5]. For the random data the distribution of error positions within the frame is uniform, but the structured data clearly shows that certain payload octets display significantly higher error-rates. It is important to note that the errors occur uniformly across the whole packet and that there are no correlations evident between the positions of errors within the frame. We interpret this result as confirming that errors are highly localised within a frame and from this we are able to assume that the error-inducing events occur over small (bit-time) time scales.

Figures 1(a) and 1(b) show bit error rate and packet error rate respectively, for a range of received optical power. The powers in these two figures differ due to the different experimental setups used. We see that these different testframes lead to substantially different BER performance. Importantly the relationship between the test data and BER results has little relationship with the packet-error rates for the same test data.

We have found that individual errored octets do not appear to be clustered within frames but are independent of each other. However, we are interested in whether earlier transmitted octets have an effect on the likelihood of a subsequent octet being received in error. We collect statistics on how many times each transmitted octet value is received in error, and also store the sequence of octets transmitted preceding this. The error counts are stored in 2D matrices (or histograms) of size 256×256, representing each pair of octets in the sequence leading up to the errored octet: one for the errored octet and its immediate predecessor, one for the predecessor and the octet before that, and so on. We normalise the error counts for each of these histograms by dividing by the matrix representing the frequency of occurrence of this octet sequence in the original transmitted data. We then scale each histogram matrix so that the sum of all entries in each matrix is 1.

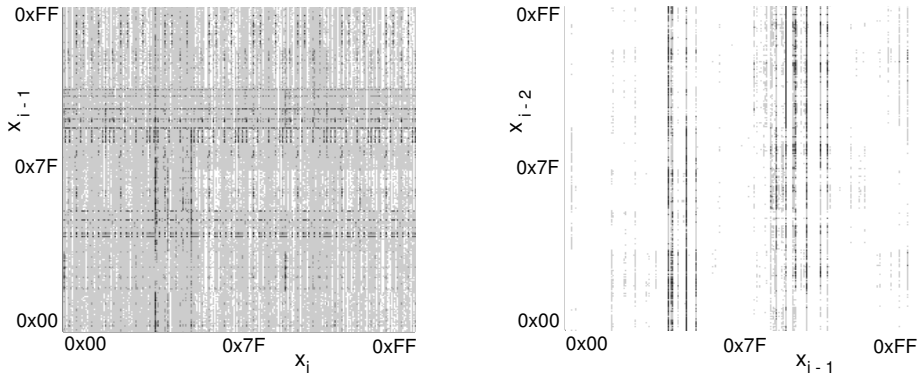
Figure 2(a) shows the error frequencies for the "current octet" X_i (the correct transmitted value of octets received in error), on the x-axis, versus the octet which was transmitted before each specific errored octet, X_{i-1}, on the y-axis. Figure 2(b) shows the preceding octet and the octet before that: X_{i-1} vs X_{i-2}.



(a) Bit error rate versus received power

(b) Packet error rate versus received power

Figure 1: Contrasting packet-error and bit-error rates versus received power



(a) Error counts for X_i vs. X_{i-1}

(b) Error counts for X_{i-1} vs. X_{i-2}

Figure 2: Error counts for pseudo-random data octets

Vertical lines in Figure 2(a) are indicative of an octet that is error-prone independently of the value of the previous octet. In contrast, horizontal bands indicate a correlation of errors with the value of the previous octet. It can be seen from Figure 2 that while correlation between errors and the value in error, or the immediately previous value, are significant, beyond this there is no apparent correlation. The equivalent plot for X_{i-2}, X_{i-3} produces a featureless white square.

It is illustrative to consider the octets which are most subject to error, and the 8B/10B codes used to represent them. In the psuedo-random data, the following ten octets give the highest error probabilities (independent of the preceding octet value): 0x43, 0x8A, 0x4A, 0xCA, 0x6A, 0x0A, 0x6F, 0xEA, 0x59, 0x2A. It can be seen that these commonly end in A, and this causes the first 5 bits of the code-group to be `01010`. The octets not beginning with this sequence in general contain at least 4 alternating bits. Of the ten octets giving the lowest error probabilities (independent of previous octet), which are 0xAD, 0xED, 0x9D, 0xDD, 0x7D, 0x6D, 0xFD, 0x2D, 0x3D and 0x8D, the concluding D causes the code-groups to start with `0011`. Fast Fourier Transforms (FFTs) were generated for data sequences consisting of repeated instances of the code-groups of 8B/10B. Examining the FFTs of the code-groups for the high error octets, the peak corresponding to the base frequency (625MHz, half the line rate) is pronounced in most cases, although there is no such feature in the FFTs of the code-groups of the low error octets.

Electrically, semiconductor lasers are just simple diodes, but the interaction between electron and photon populations within the device makes the modulation response complex. A first-order representation of the laser and driver may be obtained via a pair of rate equations, one each for electrons and photons, but DFB lasers at frequencies above 1 Gbit/s really need multiple coupled equations of this sort in order to account for spatial variations within the laser, as described in Carroll *et al.* [9]. A significant range of behaviour is possible as bias, drive conditions, and physical structure vary. In general, jitter becomes much worse if laser bias is reduced below threshold, and amplitude noise in the "0" levels becomes a problem if bias is increased. With ideal bias, just at threshold, some lasers have sufficient "memory" to react to the high frequency energy in `10101010` strings, and slight eye closure may result. Modelling confirms this result. The effect is small, but enough to increase the probability of error for such a data

block. In addition, laser drive control loops, receiver timing loops, and the more sophisticated bandwidth limiting filters in receivers, could in principle be disturbed slightly by particular bit sequences, and hence give increased error rates for those sequences.

3 Implications.

In Section 2 we documented the occurrence of error *hot-spots*: data and data-sequences with a higher probability of error, resulting in packets with those payloads being discarded with a higher-than-normal probability. An analysis of the contents of *day-trace* data along with other data derived as part of our network-monitoring work allows us to conclude that in addition to (user) data-payloads the error-concentrating effects will cause a significant level of loss due to the network and transport-layer header contents. In one hypothetical case, if a user was on a machine with an IP address that consisted of several high-error-rate octets their data could potentially be up to 100 times more likely to be corrupted and discarded at the Gigabit Ethernet layer.

In addition to increasing the chances of frame-discard due to data-contents, the occurrence of such *hot-spots* also has implications for higher level network protocols. Frame-integrity checks, such as a cyclic redundancy check, assume that there will be a uniformity of errors within the frame, justifying detection of single-bit errors with a given precision. While Jain [10] demonstrates that the FRC as used in Ethernet is sufficiently strong as to detect all 1, 2 and 3 bit errors for frames up to 8 KBytes in length, problems may be encountered for certain combinations of errors above this, and we note that many single-bit errors at the physical layer will translate into multiple bit errors after decoding [5]. Also, Stone *et al.* [11] discusses the impact this non-uniformity of error has for the checksum of TCP. This may call into question our assumption that only increased packet-loss will be the result of the error *hot-spots*. Instead of just lost packets, Stone *et al.* noted certain "unlucky" data would rarely have errors detected.

4 Conclusions.

We observe that the errors in Gigabit Ethernet in a low-power regime are not uniform as may be assumed. Examining the 8B/10B coding scheme, we have documented failures inducing, at best, poor performance and, at worst, undetected errors that may focus upon specific networks, applications and users. This content-specific effect is particularly insidious because it occurs without a total failure of the network.

Our prototype OPS system illustrates how future optical networks will consist of an increasingly large number of diverse elements, with greater limitations on the optical power budget. The design of these new networks must carefully consider the physical layer and its effects on higher level network protocols.

References

[1] D. McAuley, "Optical Local Area Network," in *Computer Systems: Theory, Technology and Applications*, A. Herbert and K. Spärck-Jones, Eds. Springer-Verlag, Feb. 2003.

[2] L. James, G. Roberts, M. Glick, D. McAuley, K. Williams, *et al.*, "Wavelength Striped Semi-synchronous Optical Local Area Networks," in *London Communications Symposium (LCS 2003)*, Sept. 2003.

[3] G. F. Roberts, K. A. Williams, R. V. Penty, I. H. White, *et al.*, "Monolithic 2x2 Amplifying Add/Drop Optical Switch for Data," in *29th European Conference on Optical Communication*, Sept. 2003.

[4] IEEE, "IEEE 802.3ah — Ethernet in the First Mile," 2004, standard.

[5] L. B. James, A. W. Moore, and M. Glick, "Structured Errors in Optical Gigabit Ethernet," in *Passive and Active Measurement Workshop (PAM 2004)*, Apr. 2004.

[6] A. X. Widmer and P. A. Franaszek, "A DC-Balanced, Partitioned-Block, 8B/10B Transmission Code," *IBM Journal of Research and Development*, vol. 27, no. 5, pp. 440–451, Sept. 1983.

[7] IEEE, "IEEE 802.3z — Gigabit Ethernet," 1998, standard.

[8] "tcpfire," 2003, http://www.nprobe.org/tools/.

[9] J. E. Carroll, J. Whiteaway, and R. Plumb, *Distributed Feedback Semiconductor Lasers*, ser. IEE Circuits, Devices & Systems Series. Co-published by the IEE and SPIE Press, 1998, no. 10.

[10] R. Jain, "Error Characteristics of Fiber Distributed Data Interface (FDDI)," *IEEE Transactions on Communications*, vol. 38, no. 8, pp. 1244–1252, 1990.

[11] J. Stone, M. Greenwald, C. Partridge, and J. Hughes, "Performance of Checksums and CRCs over Real Data," in *Proceedings of ACM SIGCOMM 2000*, Stockholm, Sweden, Aug. 2000.